Chapter 7
Scaling, Linking, and Producing Scale Scores

Introduction

A number of changes occurred as Kentucky transitioned from KIRIS to the Kentucky Commonwealth Accountability Testing System (CATS) in 1999. These changes necessitated designing and carrying out special scaling and linking procedures, so that assessment results could be linked to the existing KIRIS scales in order that data could continue to be reported on scales similar to those used in 1998 and earlier assessment years. The main changes of a technical nature resulted from 1) the decision to report results on a joint scale using both the multiple-choice (MC) and the open-response (OR) items,[1] 2) a revision of the core content that defines the test blueprint and results in changes in test content and coverage, and 3) test length became limited in select areas. The joint scale for all grades and content areas has a low of 325 and a high of 800. This was done to accommodate the multiple forms. An additional change was a decision to use reporting scales with means of 500 and standard deviations of 50 rather than using the "theta" metric having a mean of zero and standard deviation of one[2]. Transforming to this new metric will not result in scales having means and standard deviations of exactly 500 and 50 because of growth since the scales were originally developed.

The three main sections of this chapter describe the details of how the scaling and linking were accomplished and scoring tables produced. The first describes the rescaling of the 1998 data and linking to the existing scale for that year. The second provides details of procedures for scaling and linking the 1999 Kentucky Core Content Tests to those scales. The final section describes how the scoring tables were produced.

Scaling and linking was accomplished using the PARDUX and FLUX computer programs. This software was developed at CTB/McGraw-Hill to enable scaling and linking of complex assessment data such as that produced for the Kentucky Core Content Tests.

PARDUX is designed to produce a single scale by jointly analyzing data resulting from students' responses to both multiple-choice and open-response items. In PARDUX, items are calibrated based on item response theory (IRT), using the three-parameter logistic model (3PL, Lord and Novick, 1968) for multiple-choice items and the two-parameter partial credit model (2PPC, Yen, 1990) for open-response items. PARDUX is also used to link the scales developed by two calibrations through the common-item procedure developed by Stocking and Lord (1983).

[1] There were additional and significant changes in test design, policy and administration procedures that made equating impossible and therefore a link was done between KIRIS and Kentucky Core Content Tests (see Linn (1993)). "

[2] The "theta" metric had a mean of zero and a standard deviation of one in 1993, but 1994 through 1998 data reflected change over time (reference Cycle 3 Accountability Technical Report).

The FLUX program is complementary to PARDUX and allows for various linking procedures including the common population equipercentile equating method and a linear approximation to that method. The latter method was used to link the joint multiple-choice – open-response scales derived for the 1998 data, through use of PARDUX, to the existing open-response-based scales used in 1998 reporting, and based on different scaling software.

As a quality control step, all analyses were carried out by CTB/McGraw-Hill research scientists and duplicated by HumRRO scientists.

Rescaling 1998 Data And Linking To 1998 Scales

The Kentucky Core Content Tests comprise eighteen grade-by-subject tests.[3] Of these, seven involved subjects for which new scales were developed in 1999. The Arts and Humanities (A&H: grades 5, 8, and 11) and Practical Living/Vocational Studies (PL/VS: grades 5, 8, and 10) tests were not scaled previously using IRT methodology. In addition the reading assessment was moved from the 11th grade to the 10th. Hence linking was necessary only for the remaining 11 subject/grade tests. The procedures described herein, therefore, were used for the following 11 assessments:

- Mathematics at grades 5, 8, and 11,
- Reading at grades 4 and 7,
- Science at grades 4, 7, and 11, and
- Social Studies at grades 5, 8, and 11.

The 1998 assessment data in these subjects and grades were reported on scales originally developed using only the open-response items. We will refer to these scales as the 98R (1998 Reporting) scales, which were in a theta metric. The following steps were necessary to enable reporting based on all (multiple choice and open response) items on a single scale in a metric linked to the 98R scale. All these analyses were carried out using the existing 1998 students' item response data.

- Transform the 98R scale to the new reporting metric by multiplying by 50 and adding 500 (we refer to the transformed scale as the 98T scale). Compile a frequency distribution (FD) of scale scores in this metric and save it in a file.

- Calibrate all multiple-choice and open-response items jointly, resulting in item parameter estimates for all items on the 98I (1998 Initial) scale in a theta metric.

[3] Writing on-demand is a component of the Kentucky Core Content Tests, but is scored against standards established through the Kentucky Writing Portfolio process and therefore "equating" or "linking" as such was not applicable.

- Using the parameter estimates for only the open-response items, estimate scale scores for all students on the 98I scale. Compile a FD of these scale scores and save it in a file.

- Using frequency distributions of students on the 98T scale and the 98I scale, link the latter to the former by a linear approximation to equipercentile equating.

- Use the resulting multiplicative (M1) and additive (M2) constants for the linear transformation of the new parameter estimates of all items from the 98I scale to the 98T scale.

These steps were carried out for the 11 scales mentioned above. Details of the steps are presented in the following sections.

Formation of 1998 Frequency Distribution in New Metric

Each student's scale score in the 1998 reporting metric was transformed to the new (98T) metric by multiplying it by 50 and adding 500. An ungrouped frequency distribution of these scores was compiled for use in equating. This procedure was carried out for each of the 11 grade/subject combinations listed above. Assuming the original 1993 scale had a mean of 0 and standard deviation of 1, multiplying by 50 and adding 500 results in a scale having mean of 500 and standard deviation of 50. The 1998 Kentucky scales actually have different means and standard deviations due to student growth, cohort effects, etc. between 1993 and 1998.

Joint Calibration of 1998 Open-Response and Multiple-Choice Items

All 1998 items (open response and multiple choice) were calibrated using PARDUX. Very specific details of procedures used and quality control checks carried out are available in the 1998 Kentucky Scaling Specifications document. The pretest items and multiple-choice items designated as "not to be used" (having an X in the answer key field of the item documentation) were not included. The result is a set of parameter estimates for each item on a scale in the theta (98I) metric. That is, they are not yet transformed to the 98T scale. These estimates were saved in a file for use in later steps.

Estimation of Students' Scale Scores on the 98I Scale

The purpose of this step was to produce a distribution of students' scale scores based on the 98I scale but estimated using only the open-response items. This distribution is used in linking back to the 98T scale for each of the 11 assessment grades and subjects.

The file of item parameter estimates produced in the previous step was read into PARDUX along with the students' item response data. The PARDUX feature of designating items as "not to be used" was used for all multiple-choice items so that students' scale scores would be estimated from the open-response items only, as required for the equating step.

An ungrouped frequency distribution of students' scale scores estimated on the 98I scale was produced and saved in a file for use in equating.

Equating the 98I Scale to the 98T Scale

The two frequency distributions (98T and 98I) were input to the FLUX computer program. The program was then used to determine the multiplicative and additive constants necessary to transform the 98I distribution to the same metric as the 98T distribution, using a linear approximation to equipercentile equating. The result is the multiplicative (M1) and additive (M2) linear transformation constants for each subject tested in each grade. These constants are shown in Table 7.1.

Table 7.1
1998 Linear Scale Transformation Constants

Grade	Subject	M1	M2
4	Reading	31.45	548.373
	Science	27.70	541.041
5	Mathematics	32.95	551.419
	Social Studies	31.90	536.213
7	Reading	31.80	514.415
	Science	26.15	500.612
8	Mathematics	31.35	530.411
	Social Studies	36.35	509.674
11	Mathematics	35.25	531.506
	Science	27.50	542.305
	Social Studies	41.95	548.620

Scaling And Equating 1999 Kentucky Core Content Tests To 1998 Scales

In this section, we describe the procedures used to calibrate the 1999 Kentucky Core Content Tests items and transform the scales to a metric equated to that used in the 1998 reports (but in the 98T metric: mean 500 standard deviation 50). Because there was no linking back to previous scales in 7 grade/subject combinations but there was such linking in the other 11, the procedures are described separately for the two types of analyses.

Item Calibration Samples for All Grades / Subjects

In order to meet reporting deadlines, the 1999 items were calibrated before item response data were available for all students for whom reports were to be generated. When sufficient[4] data were available on all items for each subject/grade, scaling was carried out using the PARDUX computer program. Table 7.2 displays the numbers of students used for the calibrations by grade and assessment subject.

Table 7.2
Numbers of Students in Calibration Datasets

Grade	Mathematics	Reading	Science	Social Studies	Arts & Humanities	Practical Living/ Vocational Studies
4		47,645	47,637			
5	46,001			45,989	46,011	46,011
7		47,252	47,227			
8	47,931			47,894	48,028	48,020
10		44,673				44,900
11	39,801		39,741	39,607	39,956	

[4] Sufficient was defined as the availability of data from all forms of the assessment administered and generally resulted in the use of about 90 to 95% of the student data.

Calibration and Equating Procedures: Grades/Subjects Equated to 1998 Scales

Scaling and equating of these 11 grade/subject assessments was carried out using the PARDUX computer program. The equating method was based on a common set of items referred to as the anchor items, using the method derived by Stocking and Lord (1983). As mentioned above, the decision was made that only multiple-choice items would be used in the anchor set. Furthermore, the anchor items were all included on one of the six forms in each grade/subject of the assessment. Hence the anchor items were the multiple-choice items included on the linking form.

The steps used were:

- create a file of anchor parameter estimates,
- calibrate the 1999 item response data using PARDUX, and
- calculate the Stocking-Lord transformation constants.

A description of each of these steps follows.

As a first step, the parameter estimates in the 98I metric for the anchor items were selected from the file of all parameter estimates and saved in a separate file. Secondly, these estimates were transformed to the 98T metric using the constants listed in Table 7.1 and saved as an anchor file.

In the second step, the 1999 student item response data were calibrated using PARDUX. The resulting parameter estimates, including new estimates for the anchor items, were initially in a theta metric. We will designate this metric as 99I.

The Stocking-Lord procedure was then applied to the two sets of estimates and the multiplicative (M1) and additive (M2) constants were determined that would linearly transform the 99I metric to the 98T metric. These constants were then used to produce reporting results in the final scale metric. We will refer to this reporting scale as the 99R scale. The transformation constants are displayed in Table 7.3.

Calibration and Equating Procedures: Grades/Subjects Not Equated to 1998 Scales

For these grades/subjects, there was no 1998 scale, so the production of anchor files and linking step were not necessary. All items were calibrated on a scale we will also designate as the 99I scale. Transformation constants to create a final scale in the 99R reporting metric are simply a multiplicative constant, M1 of 50 and an additive constant, M2, of 500. These constants are shown in Table 7.3.

Table 7.3
1999 Linear Scale Transformation Constants

Grade	Subject	M1	M2
4	Reading	33.36	545.54
	Science	27.75	539.77
5	Arts & Humanities	50	500
	Mathematics	35.33	553.01
	Practical Living/Vocational Studies	50	500
	Social Studies	31.61	537.52
7	Reading	31.34	511.37
	Science	26.40	499.30
8	Arts & Humanities	50	500
	Mathematics	33.91	527.60
	Practical Living/Vocational Studies	50	500
	Social Studies	38.38	506.43
10	Practical Living/Vocational Studies	50	500
	Reading	50	500
11	Arts & Humanities	50	500
	Mathematics	39.85	529.85
	Science	31.11	539.99
	Social Studies	44.41	543.55

:

Producing The Scoring Tables

For each of the 18 grade/subject combinations, tables that show the corresponding scale score for each weighted (open-response and multiple-choice items received weights of 2 and 1, respectively) raw score on each form were produced. Typically, there were six forms for each grade/subject combination except in the arts and humanities and practical living/vocational studies in which there were twelve. For some forms, however, there were differences in one item between the A and B subforms. For those forms, separate tables were computed for the subforms. The procedures for computing the values in the scoring tables are specified in the document, "Computing the Raw Score to Scale Score Conversion Tables For the Kentucky Core Content Tests."

The following steps were required to produce each scoring table.

- the estimates of the parameters for the items on the form were selected from the file of estimates of all items in the grade/subject combination,

- a control file for the FLUX program was constructed specifying the M1 and M2 constants for the grade/subject,

- the FLUX program was started and the control file read in,

- the file of parameter estimates for the form was read in,

- the option to weight the open-response items by two was selected and the total weighted score specified as 72 for most subjects but 24 for A&H and PL/VS, and

- the weighted scoring table was generated and saved as a text file.

Students' score reports were produced using the values in the scoring tables which contain the scale score equivalent to each raw score and its estimated standard error.

Weighting of Raw Scores

The Kentucky Department of Education instructed CTB to differentially weight the open-response and multiple-choice items. To do this, CTB differentially weighted these items when scoring tables were produced. The computation of these tables is based on the test characteristic function (TCF, sometimes referred to as the expected score function, ESF) in IRT scaling. This function describes the relationship between the proficiency variable (in scale score units) and the expected raw score. In particular, it is derived such that the expected raw score of an individual can be determined from his/her scale score. Note that the scoring table is designed to yield the inverse, an expected scale score from an observed raw score. This is discussed further below.

The expected score function for a single multiple-choice item is simply the item response function:

$$E(r_j|\theta) = P_{j1}(\theta) \qquad [1]$$

where $E(r_j|\theta)$ represents the expected raw score on item *j* given the scale score, θ, and $P_{j1}(\theta)$ is the probability of a correct score (score of *1*) given a scale score of θ. Given a student's scale score, the function provides the probability of a correct response, which is the expected score on the item for a student having that scale score.

For a test comprised of *n* multiple-choice items, the TCF is the sum of the ESFs of the *n* items:

$$\zeta(\theta) = \sum_{j=1}^{n} E(r_j|\theta) = \sum_{j=1}^{n} P_{j1}(\theta) \qquad [2]$$

and it represents the relationship between the expected number of correct responses on the *n* items and students' scale scores.

For an open-response item scored on an m_j-point scale (*0* to *m-1*), the ESF is given by

$$E(r_j|\theta) = \sum_{k=1}^{m_j-1} kP_{jk}(\theta) \qquad [3]$$

where $P_{jk}(\theta)$ represents the probability of a student with scale score θ getting a score of *k* $(k = 1,2,\ldots,m_j - 1)$ on item *j*. Strictly speaking, we could write [3] as summing from the lowest score, *0*, to the highest score, $m_j - 1$, but note that the term in the expression for *k=0* is zero, so that term is unnecessary.

Note that expression [1] can be considered to be a special case of [3] in which $m_j = 2$ because in this case (a multiple-choice item scored *1* or *0*)

$$E(r_j|\theta) = \sum_{k=1}^{m_j-1} kP_{jk}(\theta) = \sum_{k=1}^{1} kP_{jk}(\theta) = 1 \times P_{j1}(\theta) = P_{j1}(\theta) \qquad [4]$$

which is identical to [1]. Hence the expression in [3] may be used for either multiple-choice or open-response items.

The test characteristic function for a mixture of multiple-choice and open-response items on an n-item test thus can be written as

$$\zeta(\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j - 1} k P_{jk}(\theta) \qquad [5]$$

The expression for the probabilities is somewhat complex. For multiple-choice items it may be written as

$$P_{j1}(\theta) = P(x_j = 1 | \theta,\ a_j,\ b_j,\ c_j) = c_j + (1 - c_j) \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}} \qquad [6]$$

$$(j = 1, 2, \dots, n)$$

In this model:

$P_{j1}(\theta)$ is the probability of a response of *1* given $\theta,\ a_j,\ b_j,\ c_j$, where the "*1*" (second) subscript on P indicates specifically that we are dealing with the probability of response category *1*,

x_j is the response to the *j*th item of an *n*-item instrument,

θ is the proficiency variable,

a_j is a discrimination parameter of the *j*th item,

b_j is a difficulty or location parameter of the *j*th item,

c_j is the lower asymptote of the ICC of the *j*th item, and

Note that for any dichotomously-scored (two score points) item such as the multiple-choice items under consideration here, there are two possible outcomes, correct and incorrect. In our notation we denote a correct item score as a "*1*" and an incorrect score as a "*0*". The probability of an incorrect score is simply one minus the probability of a correct score,

$$P_{j0}(\theta) = 1 - P_{j1}(\theta)$$

and we need not represent that probability in our model.

For the open-response items the probabilities of the *k* responses are given by

$$P_{jk}(\theta) = \frac{e^{\sum_{i=1}^{k}(\alpha_j\theta - \gamma_{ji})}}{\sum_{t=0}^{m_j - 1} e^{\sum_{i=1}^{t} a_j(\theta - \gamma_{ji})}} \qquad [7]$$

$$\gamma_0 = 0$$

$$(k = 0,1,2,\ldots,m_j - 1)$$

$$(j = 1,2,3,\ldots,n)$$

where there is one α parameter for each item and a γ parameter for each response category except zero, for each item. Because each examinee must receive one of the m_j category scores, the probabilities for a given value of θ sum to *1.0* so the probability for category *0* is simply

$$P_{j0}(\theta) = 1.0 - \sum_{k=1}^{m_{j-1}} P_{jk}(\theta) \qquad [8]$$

Computing the values necessary to create a raw score to scale score table involves inverting the function in [5] so the expected raw score could be computed as a function of the scale score.

Given the complexity of the expressions for the probabilities, the inverse function is extremely complex and requires a numerical method to perform the estimation.

Weighting Sets of Items

For the Kentucky Core Content Tests, the decision has been made that open-response items should receive twice the weight of multiple-choice items in determining student performance. This weighting can be accomplished by inserting a weighting factor into equation [5]. The new equation reflecting the weights is

$$\zeta(\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j - 1} w_j k P_{jk}(\theta) \qquad [9]$$

$$\left\{ \begin{array}{l} w_j = 1\, for\, multiple - choice\ items \\ w_j = 2\, for\, open - response\ items \end{array} \right\}$$

where *w* is the weighting factor (2 for the KCCT).